UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/649,909	08/26/2003	Satyanarayana Dharanipragada	YOR920030259US1	5755

48813 7590 10/03/2008

LAW OFFICE OF IDO TUCHMAN (YOR)
ECM #72212
PO Box 4668
New York, NY 10163-4668

EXAMINER	
COLUCCI, MICHAEL C	
ART UNIT	PAPER NUMBER
2626	

NOTIFICATION DATE	DELIVERY MODE
10/03/2008	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

pair@tuchmanlaw.com
idotuchman@gmail.com

Office Action Summary	Application No.	Applicant(s)	
	10/649,909	DHARANIPRAGADA ET AL.	
	Examiner	Art Unit	
	MICHAEL C. COLUCCI	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE *3* MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☐ Responsive to communication(s) filed on _____.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-27* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-27* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *26 August 2003* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☒ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☒ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☐ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____.
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

Response to Arguments

1. Applicant's arguments, see Remarks, pages 19-21, filed 07/07/2008, with respect to the rejection(s) of claim(s) 1 and 6 under 35 U.S.C. 103(a) have been fully considered and are persuasive. Therefore, the rejection has been withdrawn. However, upon further consideration, a new ground(s) of rejection is made in view of Yang US 20010010039 A1 (hereinafter Yang). Examiner concurs that the reference of Chandrasekar 6,578,032 is not relevant to the current scope of the present invention relevant to multiple speakers, training, and phonemic modeling. Though Chandrasekar teaches comparison of clusters, the clustering is unrelated to the present invention of speaker classification and modeling. Therefore the reference of Chandrasekar has been withdrawn. Additionally, the reference of Newman et al. US 6,151,575 (hereinafter Newman) was withdrawn, as Examiner believes it is redundant with respect to the aforementioned *Yang*.

Claim Rejections - 35 USC § 103

2. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

3. Claims 1, 5, 6, 10, 11, 15, and16 are rejected under 35 U.S.C. 103(a) as being unpatentable over Chang et al. US 6567776 B1 (hereinafter Chang) in view of Yang US 20010010039 A1 (hereinafter Yang).

Re claims 1, 6, 11, and 16, Chang teaches a method for generating speech recognition models, the method comprising:

converting speech spoken from a plurality of female speakers (Col. 1 lines 15-49) into a first set of recorded phonemes training data (Col. 5 line 45 – Col. 6 line 67);

converting speech spoken from a plurality of male speakers (Col. 1 lines 15-49) into a second set of recorded phonemes training data (Col. 5 line 45 – Col. 6 line 67);

receiving a first speech recognition model based on the first set of recorded phonemes training data (Col. 5 line 45 – Col. 6 line 67);

receiving a second speech recognition model based on the second set of recorded phonemes training data (Col. 5 line 45 – Col. 6 line 67);

determining a difference in model information between the first speech recognition model and the second speech recognition model (Col. 5 line 45 – Col. 6 line 67);

However, Chang fails to teach phoneme training data

creating a gender-independent speech recognition model based on the first set of recorded phonemes training data and the second set of recorded phonemes training data if the difference in model information is insignificant.

Yang teaches very well known techniques of speech recognition, wherein difference are evaluated between all voice types, wherein Yang teaches human speech

is generated according to a shape of vocal tract and its temporal transition. The shape of vocal tract, which depends on the shape or size of the vocal organ, inevitably shows individual differences. On the other hand, the pattern of time sequence of the vocal tract, which also depends on an uttered word that, shows a small individual difference. Therefore, features of utterance should be divided into two factors: the shape of the vocal tract and its temporal pattern. The former shows large difference from speaker to speaker whereas the latter one shows small difference. So if the difference based on the shape of the vocal tract is somehow normalized, the speech of specified speakers can be recognized using only the utterances of a small number of speakers. The difference in the shape of the vocal tracts causes different frequency spectra. One of the methods to normalize the spectral difference among speakers is to classify voice input by matching it with phoneme templates which are made for unspecified speakers. This operation provides similarity, which does not depend very much on the differences among speakers. Meanwhile, the temporal pattern of vocal tract is considered to have small individual difference (Yang [0004]).

Further, Yang teaches speech recognition method comprises the step of training a Phoneme Similarity Vector (PSV) model on the initial part to create an initial part model having trained initial part model parameters, the step of training a PSV on the final part to create a final part model having trained final part model parameter, the step of training a PSV on the training speech syllable to create a syllable model using the trained initial part parameter values and the trained final part parameter values as starting parameters for the syllable model, the step of operating on an object speech

sample with the syllable model, the step of recognizing the object speech sample as an object speech syllable based on a degree of match of the object speech sample to the syllable model, and the step of representing the object speech sample as a Chinese character in accordance with the object speech syllable (Yang [0014]).

Furthermore, with respect to distance comparison, Yang teaches a user creating a speech signal to accomplish a given task. In the second step, the spoken output is first recognized in that the speech signal is decoded into a series of phonemes that are meaningful according to the phoneme templates. The acoustic analysis portion 30 analyses speech inputs and the extracted LPC (Linear Predictive Coding) cepstrum coefficients and delta power. The extracted parameters are matched with many kinds of phoneme templates, and static phoneme similarity and the first order regression coefficients of phoneme similarity are calculated in the similarity calculation portion 40. After that, the time sequence of those number of phoneme templates to define a dimensional similarity coefficient vectors and regression coefficient vectors can be obtained. In the similarity calculation portion 40, mahalanobis' distance algorithm is employed for distance measure, where covariance matrixes for all of the phonemes are assumed to be the same. The meaning of the recognized words is obtained by the post processor that uses a dynamic programming to match inputted word with the real word and the word having been previously recognized by phoneme similarity calculation (Yang [0036]).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Chang to incorporate phoneme training

data and creating a gender-independent speech recognition model based on the first set of recorded phonemes training data and the second set of recorded phonemes training data if the difference in model information is insignificant as taught by Yang to allow for the acquisition of various speech parameters from multiple speakers where phoneme templates are made for unspecified speakers, wherein temporal patterns and frequency spectra are analyzed to find the difference between speakers based on a vocal tract (i.e. a male and female can have different voice features) (Yang [0004]).

Re claims 5, 10, and 15, Chang teaches method of claim 1, wherein the first speech recognition model, second speech recognition model, and gender-independent speech recognition model (Col. 5 line 45 – Col. 6 line 67) are Gaussian mixture models .

However, Chang fails to teach speech recognition models that are Gaussian mixture models.

Yang teaches the use of the continuous mixture Gaussian density models. With these methods, spectral parameters are used in speech recognition as a feature parameter and an enormous number of speakers are generally required for training. It also costs very large memory in order to get high recognition rate. If the standard patterns for speaker independent speech recognition can be produced from a small number of speakers, the size of computation will be much smaller than usual. Therefore, human power and computation are saved and speech recognition technique can be easily handled to various applications. For the purpose mentioned above, we proposed our invention of speech recognition apparatus using the similarity vectors as

feature parameters. In this method, word templates trained with a small number of speakers yield high recognition rates in speaker-independent recognition. To realize the speech recognition technology in real applications, speech recognizer must be robust to noisy environments and spot intended words from background noise and unintended utterances. Furthermore, speech recognizer must retain high quality performance on portable devices. For these reasons, our invention was focused on small-size programming code but high accuracy rate for portable device which can be built-in a Chinese speech recognition system (Yang [0007]).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Chang to incorporate gender-independent speech recognition model that are Gaussian mixture models as taught by Yang to allow for a less costly approach that produces higher accuracy for a speech recognition system, wherein recognition rates are based on speaker-independent recognition and modeling (Yang [0007]).

4. Claims 2-4, 7-9, and 12-14 are rejected under 35 U.S.C. 103(a) as being unpatentable over Chang et al. 6567776 (hereinafter Chang) in view of Yang US 20010010039 A1 (hereinafter Yang) and further in view of Kanevsky et al. US 6529902 (hereinafter Kanevsky).

Re claims 2, 7, and 12, Chang in view of Yang fails to teach the method of claim 1, wherein whether the model information is insignificant is based on a threshold model quantity.

Kanevsky teaches the Kullback-Leibler distance between any two topics is at least h, where h is some sufficiently large threshold (Kanevsky Col. 5, lines 9-11). Further, Kanevsky teaches using Kullback-Leibler distance, one can check which pairs of topics are sufficiently separated from each other. Topics that are close in this metric could be combined together (Kanevsky Col. 12, lines 44-47).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Chang in view of Yang to incorporate the model information is insignificant is based on a threshold model quantity as taught by Kanevsky to allow for an improved language modeling for off-line automatic speech decoding and machine translation (Kanevsky Col. 2, lines 50-52).

Re claims 3, 8, and 13, Chang in view of Yang fails to teach the method of claim 1, wherein determining the difference in model information includes calculating a Kullback Leibler distance between the first speech recognition model and second speech recognition model.

Kanevsky et al. teaches that for two different sets, one can define a Kullback-Leibler distance using the frequencies of the sets. [With the distance] one can check which pairs of topics are sufficiently separated from each other. Topics that are close in this metric could be combined together (Kanevsky Col. 12, lines 42-47).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Chang in view of Yang to incorporate the determining the difference in model information includes calculating a Kullback Leibler

distance between the first speech recognition model and second speech recognition model as taught by Kanevsky to allow for an improved language modeling for off-line automatic speech decoding and machine translation (Kanevsky Col. 2, lines 50-52).

Re claims 4, 9, and 14, Chang in view of Yang fails to teach the method of claim 3, wherein whether the model information is insignificant is based on a threshold Kullback Leibler distance quantity.

Kanevsky teaches the Kullback-Leibler distance (Kanevsky Col. 5, lines 9-11) between any two topics is at least h, where h ~s some sufficiently large threshold, also they teach (Kanevsky Col. 12, lines 44-47) that while using the Kullback-Leibler distance, one can check which pairs of topics are sufficiently separated from each other, and that topics that are close in this metric could be combined together).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Chang in view of Yang to incorporate whether the model information is insignificant is based on a threshold Kullback Leibler distance quantity as taught by Kanevsky to allow for an improved language modeling for off-line automatic speech decoding and machine translation, wherein a sufficiently large threshold indicates separate or combinational probabilities (Kanevsky Col. 2, lines 50-52).

5. **Claims 17-27 are rejected under 35 U.S.C. 103(a) as being unpatentable over Wark US 20030231775 (hereinafter Wark) in view of Chang et al. 6567776**

(hereinafter Chang) and further in view of Yang US 20010010039 A1 (hereinafter Yang).

Re claims 17, 21, and 24, Wark teaches a system for recognizing speech data from an audio stream originating from one of a plurality of data classes ([0094]) system comprising:

a computer processor;

a receiving module configured to receive a current feature vector of the audio stream ([0094]);

a first computing module configured to compute a current vector probability ([006]) that the current feature vector belongs to one of the plurality of data classes ([0094]);

a second computing module configured to compute an accumulated confidence level that the audio stream belongs to one of the plurality of data classes based on the current vector probability ([0060]) and on previous vector probabilities ([0146] & Fig. 4, adjacent, previous and current segment/frame);

a weighing module ([0142]) configured to weigh class models based on the accumulated confidence ([0146]); and

a recognizing module configured to recognize the current feature vector ([0094]) based on the weighted class models ([0130]); and

However, Wark in view of Chang fails to teach a plurality of data classes that include a first speech recognition model based on recorded phonemes originating from

a first set of speakers, a second speech recognition model based on recorded phonemes from a second set of speakers, and a third speech recognition model based on recorded phonemes originating from both the first and second set of speakers having insignificant differences in information.

Chang teaches that it is well known in related art, we learn that speaker cluster models have been applied to speaker-independent speech recognition and speaker adaptation. Although used in different application fields, the speaker cluster models are built in the same training phases. A training phase starts with dividing speakers into different speaker clusters. Then a cluster-dependent model is independently trained for each speaker cluster by using the speech data of the speakers belonging to the cluster. The collection of all cluster-dependent models then forms a speaker cluster model. Most approaches in building speaker cluster models are focused on means of dividing speakers into clusters, especially in finding measurement of similarities across speakers. Some speaker clustering methods reported in articles of the related art are as follows: 1. Using acoustic distances across speakers to measure similarities across speakers (Chang Col. 1 lines 15-49).

Further, Chang teaches speaker based modeling representing in a tree form for purposes of explanation, wherein in the first level (root) of the tree we use all of the speech data to train a speaker-independent model. All speakers are then clustered according to gender. They are clustered into the male speaker cluster 102 and female speaker cluster 104 to train a gender-dependent model. This is the second level of the tree. Finally, the speakers within each gender group are clustered into two speaker

clusters. For example, the male speaker cluster 102 is clustered into the speaker clusters M1112 and M2114, respectively. The female speaker cluster 104 is clustered into the speaker clusters F1122 and F2124, respectively. Hence, the third level of the tree has four clusters. In this step, we use acoustic distances across speakers to measure similarities across speakers. (Chang Col. 4 line 56 – Col. 5 line 25).

Furthermore, Chang teaches a speaker-independent model, which is built using maximum likelihood as the training criteria and is the first level (cluster 100) of the speaker cluster model, to recognize the speech signal. Its result is used for comparing with the results of other experiments. Because this level only comprises one speaker cluster, the result is the same regardless the value of .xi.. B. Further adjust the parameters of the model used in experiment A (cluster 100) using the discriminative training method. It is shown in Table 1 that a better recognition result is achieved using the discriminative training method. Because the training method of the speaker cluster model introduced by the present invention uses the discriminative training method, the recognition model used for comparison is also established by using the discriminative training method. However, the discriminant function g.sub.i of the present invention is different from the discriminant function h.sub.i of the related art. C. Perform discriminative training on the gender dependent model (male speaker cluster 102 and female speaker cluster 104) in. the second level of the tree. Because speakers of different gender clusters have very different characteristics, we will not adjust parameters across different gender clusters. That means that the discriminative training performed on the parameters of the male speaker cluster 102 only uses speech data

uttered by male speakers. The discriminative training performed on the parameters of the female speaker cluster 104 only uses speech data uttered by female speakers. It is shown in Table 1 that the recognition result using the gender-dependent model is superior to that using the speaker-independent model. Because the gender-dependent model is a simple plain-structured speaker cluster model, the speaker cluster model can readily manage recognition problems caused by differences between speaker characteristics, improving the recognition result of speaker-independent speech recognition (Chang Col. 5 line 45 – Col. 6 line 67).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Wark to incorporate a plurality of data classes that include a first speech recognition model based on recorded phonemes originating from a first set of speakers, a second speech recognition model based on recorded phonemes from a second set of speakers as taught by Chang to allow for the training of a speaker independent model based on gender dependent models, wherein recognition results are improved where problems due to differences in speaker characteristics are minimized to enhance modeling and training (Chang Col. 5 line 45 – Col. 6 line 67).

However, Wark in view of Chang fails to teach phoneme training data creating a gender-independent speech recognition model based on the first set of recorded phonemes training data and the second set of recorded phonemes training data if the difference in model information is insignificant.

Yang teaches very well known techniques of speech recognition, wherein difference are evaluated between all voice types, wherein Yang teaches human speech is generated according to a shape of vocal tract and its temporal transition. The shape of vocal tract, which depends on the shape or size of the vocal organ, inevitably shows individual differences. On the other hand, the pattern of time sequence of the vocal tract, which also depends on an uttered word that, shows a small individual difference. Therefore, features of utterance should be divided into two factors: the shape of the vocal tract and its temporal pattern. The former shows large difference from speaker to speaker whereas the latter one shows small difference. So if the difference based on the shape of the vocal tract is somehow normalized, the speech of specified speakers can be recognized using only the utterances of a small number of speakers. The difference in the shape of the vocal tracts causes different frequency spectra. One of the methods to normalize the spectral difference among speakers is to classify voice input by matching it with phoneme templates which are made for unspecified speakers. This operation provides similarity, which does not depend very much on the differences among speakers. Meanwhile, the temporal pattern of vocal tract is considered to have small individual difference (Yang [0004]).

Further, Yang teaches speech recognition method comprises the step of training a Phoneme Similarity Vector (PSV) model on the initial part to create an initial part model having trained initial part model parameters, the step of training a PSV on the final part to create a final part model having trained final part model parameter, the step of training a PSV on the training speech syllable to create a syllable model using the

trained initial part parameter values and the trained final part parameter values as starting parameters for the syllable model, the step of operating on an object speech sample with the syllable model, the step of recognizing the object speech sample as an object speech syllable based on a degree of match of the object speech sample to the syllable model, and the step of representing the object speech sample as a Chinese character in accordance with the object speech syllable (Yang [0014]).

Furthermore, with respect to distance comparison, Yang teaches a user creating a speech signal to accomplish a given task. In the second step, the spoken output is first recognized in that the speech signal is decoded into a series of phonemes that are meaningful according to the phoneme templates. The acoustic analysis portion 30 analyses speech inputs and the extracted LPC (Linear Predictive Coding) cepstrum coefficients and delta power. The extracted parameters are matched with many kinds of phoneme templates, and static phoneme similarity and the first order regression coefficients of phoneme similarity are calculated in the similarity calculation portion 40. After that, the time sequence of those number of phoneme templates to define a dimensional similarity coefficient vectors and regression coefficient vectors can be obtained. In the similarity calculation portion 40, mahalanobis' distance algorithm is employed for distance measure, where covariance matrixes for all of the phonemes are assumed to be the same. The meaning of the recognized words is obtained by the post processor that uses a dynamic programming to match inputted word with the real word and the word having been previously recognized by phoneme similarity calculation (Yang [0036]).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Chang to incorporate phoneme training data and creating a gender-independent speech recognition model based on the first set of recorded phonemes training data and the second set of recorded phonemes training data if the difference in model information is insignificant as taught by Yang to allow for the acquisition of various speech parameters from multiple speakers where phoneme templates are made for unspecified speakers, wherein temporal patterns and frequency spectra are analyzed to find the difference between speakers based on a vocal tract (i.e. a male and female can have different voice features) (Yang [0004]).

Re claims 18, 22, and 25, method of claim 17, wherein computing the current vector probability ([0060]) includes estimating a posteriori class probability for the current feature vector ([0146] & Fig. 4, adjacent, previous and current segment/frame).

Re claims 19, 23, and 26, method of claim 17, wherein computing the accumulated confidence level further comprising weighing the current vector ([0094]) probability ([0060]) more than the previous vector probabilities ([0146] & Fig. 4, adjacent, previous and current segment/frame).

Re claims 20 and 27, method of claim 17, further comprising determining if another feature vector is available for analysis ([0094]).

Conclusion

6. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 4910782 A. US 20030088414 A1, US 6813604 B1, US 6205424 B1.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Michael C. Colucci whose telephone number is (571)-270-1847. The examiner can normally be reached on 9:30 am - 6:00 pm, Monday-Friday.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on (571)-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/Michael C Colucci/
Examiner, Art Unit 2626
Patent Examiner
AU 2626
(571)-270-1847
Michael.Colucci@uspto.gov

/Richemond Dorvil/
Supervisory Patent Examiner, Art Unit 2626